ROCZNIKI PSYCHOLOGICZNE/ANNALS OF PSYCHOLOGY
2015, XVIII, 1, 9-24

ŁUKASZ BUDZICZ
Uniwersytet im. Adama Mickiewicza w Poznaniu
Instytut Psychologii

DYSKUSJA „PO STAPELU”. WOKÓŁ RZETELNOŚCI BADAŃ I PUBLIKACJI W PSYCHOLOGII

W roku 2011 wykryto oszustwo Diederika Stapela. Okazało się, że nie tylko Stapel fałszował dane, lecz także czasopisma nie dostrzegły wielu oczywistych błędów oraz zachęcały do przekłamań (np. pomijania badań z wynikami nieistotnymi). Równolegle ukazał się artykuł Simmonsa, Nelsona i Simonsohna (2011) poświęcony wątpliwym praktykom badawczym, które mogą znacząco zwiększać odsetek wyników fałszywie pozytywnych poprzez arbitralne decyzje dotyczące analizy i prezentacji danych. Niedługo potem pojawiły się wyniki badań wskazujących na to, że znaczna część badaczy przyznaje się do stosowania takich praktyk oraz że są one powszechnie akceptowane. Wydarzenia te zaowocowały szeroką dyskusją dotyczącą rzetelności danych w psychologii. Autor omawia najważniejsze punkty dyskusji, wskazując też, w jaki sposób niski stopień dojrzałości teorii, brak konsensusu o do zasad stosowania technik badawczych i interpretowania wyników oraz nierealistyczne wymagania redakcji czasopism empirycznych mogły się przyczynić do kryzysu.

Słowa kluczowe: oszustwo w psychologii, Diederik Stapel, „psychologia fałszywie pozytywna”, wybiórcze publikowanie.

Adres do korespondencji: ŁUKASZ BUDZICZ – Instytut Psychologii, Uniwersytet im. Adama Mickiewicza w Poznaniu, ul. Szamarzewskiego 89, 60-568 Poznań; e-mail: lukasz.budzicz@gmail.com

Autor dziękuje prof. Jerzemu M. Brzezińskiemu i dwóm anonimowym recenzentom za wartościowe uwagi do pierwszej wersji artykułu.

STAPEL I KRYZYS PSYCHOLOGII AD 2014

Choć dekretowanie „kryzysu" psychologii jest czymś regularnym i towarzyszy nam od bardzo dawna (Asch, 1952/1987; Ring, 1967; Elms, 1975; Bevan, 1991; Staats, 1999; Rozin, 2001; Rand i Ilardi, 2005), wydaje się, że ostatnie lata są szczególne. W relatywnie krótkim czasie ukazało się wiele tekstów poddających w wątpliwość rzetelność typowych danych w psychologii (cały numer 6 *Perspectives on Psychological Science* – vol. 7, 2012; Asendorf i in., 2013; Ferguson, 2013; John, Loewenstein i Prelec, 2012; Kepes i McDaniel, 2013; LeBel i Peters, 2011; Masicambo i Lalande, 2012; Mitchell, 2012; Murayama, Pekrun i Fiedler, 2013; Simmons, Nelson, i Simonsohn, 2011; Simonsohn, 2013; Simonsohn, Nelson i Simmons, 2014; Tressoldi, 2012; Wicherts, Bakker i Molenaar, 2011; w literaturze polskiej: Brzeziński, 2012 oraz Klebaniuk, 2012). Specjalny numer *Perspectives on Psychological Science* zatytułowano *Crisis of confidence* (Pahsler i Wagenmakers, 2012), co można wymownie przetłumaczyć jako „kryzys pewności", ale też „kryzys zaufania".

Najważniejszym powodem „wzmożenia" jest wykrycie trwającego kilkanaście lat oszustwa Stapela (większość cytowanych wyżej artykułów je przywołuje). Historia odbiła się szerokim echem wśród specjalistów (i nie tylko) i nie ma potrzeby przypominać tutaj szczegółów, które są szeroko opisane w innych miejscach (Bhattacharjee, 2013; Klebaniuk, 2012; Levelt Committee, Noort Committee, Drenth Committee, 2012; dalej: Levelt, 2012). Wystarczy tylko wspomnieć, że po dochodzeniu specjalnej komisji wycofano prawie 60 artykułów z okresu 15 lat, opublikowanych w najbardziej uznanych psychologicznych czasopismach.

Istotność historii Stapela nie polega na tym, że w psychologii znalazł się jeden oszust. Skoro są w innych dyscyplinach, trudno spodziewać się, żeby psychologia była w tym względzie jakaś szczególna. Problem w tym, że kilkadziesiąt tekstów z przeszło dwoma setkami badań nie zostało zakwestionowanych ani na etapie *peer-review*, ani – przede wszystkim – poprzez mechanizm intersubiektywnej sprawdzalności (niezależne replikacje). To znacząco odróżnia aferę Stapela od innych oszustw naukowych, zwłaszcza w naukach przyrodniczych (por. Stroebe, Postmes i Spears, 2012), w których wykrycie oszustwa często było spowodowane niemożliwością powtórzeń eksperymentów. Fizycy byli świadkami afery podobnego kalibru. Jan Hendrik Schön, swego czasu młoda gwiazda fizyki materiałowej, publikował jeden artykuł co dwa tygodnie w czasopismach pokroju *Science* i *Nature*. Żadne laboratorium nie potrafiło uzyskać materiałów podobnej jakości jak rzekomo przez niego wytwarzane, przez co zaczęto uważnie przyglądać się jego badaniom. Od pierwszych publikacji eksperci potrzebowali niecałych dwóch lat, żeby wykryć systematyczne anomalie w danych i za-

kwestionować rzetelność jego pracy (Reich, 2009). Natomiast przed wyjściem na jaw oszustwa Stapela nie pojawiła się żadna systematyczna krytyka jego odkryć oraz nie ukazała się, według mojej najlepszej wiedzy, żadna nieudana replikacja jego badań. Sam Stapel „wpadł" z powodów, jak na skalę oszustwa, wysoce prozaicznych (np. chwalił się współpracownikowi udanymi wynikami, ale nie był w stanie pokazać danych surowych; Bhattacharjee, 2013).

Samemu oszustwu daleko było do doskonałości. Dokonana *post factum* analiza danych statystycznych w artykułach wykazała szereg nieprawidłowości. Eksperci zwracali uwagę na takie nieprawdopodobne sytuacje, jak identyczne dane w niezależnych eksperymentach (np. średnie i odchylenia standardowe), niewystępowanie brakujących danych, bardzo niskie statystyki F wszędzie tam, gdzie nie spodziewano się istotnych efektów, zbyt silne efekty jak na skale o niskich rzetelnościach lub skale z jedną pozycją. Gdy analizowano dane surowe, znajdywano mechanicznie przeklejone kolumny zmiennych. W zasadzie każdy zakwestionowany artykuł miał jakieś nieprawidłowości (Levelt, 2012, s. 69-100). Fabrykowanie danych nie jest, wbrew pozorom, łatwe. To, że najlepsze czasopisma psychologiczne systematycznie przepuszczały takie dane, świadczy o tym, że albo recenzenci nie mieli dostatecznych kompetencji (przede wszystkim statystycznych), albo też przygotowywali recenzje „po łebkach". APA w standardowym ogłoszeniu o poszukiwaniu recenzentów informuje, że „recenzowanie jest czasochłonne; potrzeba około 1-4 godzin na jeden manuskrypt" (można je znaleźć w bazie PsycARTICLES po wpisaniu zapytania „reviewers wanted"). Przyjmując nawet liczbę czterech godzin za wystarczającą, trudno nie postawić retorycznego pytania, czy rzeczywiście przez cztery godziny można przeczytać z dużą uwagą artykuł opisujący 3-6 badań, przeanalizować podstawy teoretyczne, sprawdzić cytowane źródła, zastanowić się nad sensownością procedury i wykorzystanych narzędzi, nie mówiąc o skontrolowaniu danych surowych czy sprawdzeniu poprawności obliczeń (o napisaniu samej recenzji nie wspominając). Stawiałbym tezę, że przez taki czas recenzent może co najwyżej sprawdzić podstawowe standardy edytorskie, wytknąć oczywiste błędy i przedstawić w recenzji własne poglądy na sprawę.

Komitet powołany do zbadania afery nie obwinił tylko samego Stapela. Redaktorzy i recenzenci czasopism chętnie bowiem przyjmowali bez zastrzeżeń (prawie) zawsze udane badania Stapela, ale również zachęcali go oraz współautorów jego artykułów do przekłamań. Cytując raport: „[…] współautorzy [artykułów Stapela] przyznawali, że czasami redaktorzy i recenzenci prosili, żeby pewne zmienne były usunięte, ponieważ dzięki temu wyniki byłyby bardziej zgodne z rozumowaniem i tokiem narracji. Skutkowało to usuwaniem niepożądanych rezultatów. Recenzenci również wymagali, aby nie wszystkie przepro-

wadzone analizy były opisywane, przykładowo poprzez niewspominanie «niewychodzących» hipotez, które początkowo były zakładane jako prawdziwe. Czasami recenzenci nalegali na przeprowadzenie dodatkowych badań «pilotażowych», które były potem raportowane jako przeprowadzone przed właściwym badaniem. W ten sposób procedura badań oraz wybór bodźców były uzasadnione rzekomo wcześniejszą wiedzą […]" (Levelt, 2012, s. 53).

Stapel pokazał, że inteligentna osoba jest w stanie łatwiej publikować i robić karierę za pomocą „gładkich" wyników, podpartych atrakcyjną narracją, niż danych zebranych rzetelnie, które najprawdopodobniej byłyby bardziej niejednoznaczne i mniej efektowne. Jednak oszustwo Stapela może wskazywać na dużo bardziej fundamentalny problem, a mianowicie na to, że w psychologii znaczna część odkryć jest ze sobą słabo powiązana teoretycznie, ma miejsce bardziej „zbieractwo" faktów niż budowanie dobrze zintegrowanego systemu wiedzy. Nowe odkrycia trudno umieścić na tle szerszego krajobrazu teoretycznego, gdyż albo go nie ma, albo (zapewne częściej) jest zbyt ogólny, aby jednoznacznie stwierdzić, jakie powinny być wyniki badań. W jednym ze swoich sfałszowanych badań Stapel „pokazał", że w bardziej chaotycznym środowisku (np. zanieczyszczonym śmieciami) ludzie w większym stopniu dyskryminują mniejszości (np. siadają dalej od takich osób na publicznej ławce; Stapel i Lindenberg, 2011). Ta i inne jego hipotezy nie wydają się w żaden sposób szczególne. Są to badania, jakich mnóstwo w literaturze. Nie opisują jakichś przełomowych lub sensacyjnych odkryć, zwłaszcza takich, które byłyby sprzeczne z dotychczasową wiedzą (tak jak Bem, 2011 w swoim artykule o prekognicji). Trudno też wskazać jakieś (meta)teorie psychologiczne, z którymi byłyby niezgodne. Natomiast można (Stapel to zrobił) dorobić do nich sensownie brzmiące uzasadnienie teoretyczne i dodać sporo źródeł ukazujących podobne efekty. Efekty, choć przekonujące, nie są jakieś szczególnie silne (sam Stapel zwracał uwagę na to, żeby sfabrykowane efekty nie były bardzo mocne, gdyż będą mało przekonujące dla redaktorów, por. Bhattacharjee, 2013).

Oszustwo na taką skalę jest czymś niezwykle rzadkim nawet w naukach przyrodniczych (por. Stroebe i in., 2012). W psychologii jest tym bardziej szczególne, gdyż oszustwa są tu wykrywane rzadko. Zasadnicze pytanie brzmi, czy Stapel był „czarną owcą" (jak sugeruje organizacja psychologów społecznych *European Association of Social Psychology*, 2012), czy może przykładem zjawiska wcale nierzadkiego w psychologii. Definitywnej odpowiedzi nie ma, przypadek Stapela pokazuje jednak, że wskazany jest niepokój co do możliwości społeczności naukowej w wykrywaniu takich oszustw. Jednak fałszerstwo polegające na tworzeniu nieistniejących danych to niejedyny problem, który może poważnie zniekształcać obraz rzeczywistości w nauce. W dyskusjach wokół stanu

psychologii pojawiają się głosy, że wcale niemały odsetek badaczy może przekłamywać rzeczywistość w dużo bardziej subtelny sposób.

SUBTELNIEJSZE PRZEKŁAMANIA, CZYLI PSYCHOLOGIA FAŁSZYWIE POZYTYWNA

W mniej więcej tym samym czasie, gdy wybuchła afera Stapela, ukazał się artykuł Simmonsa i współautorów (2011) o „psychologii fałszywie pozytywnej” (gra słów z połączenia „błędu fałszywie pozytywnego” i „psychologii pozytywnej”; ang. *false-positive psychology*). Tekst zdobył sporą popularność (ponad 800 cytowań w Google Scholar; stan na kwiecień 2015) i spekulowałbym, że afera Stapela wydatnie się do tego przyczyniła. Nie omawia on zasadniczo nowych idei (np. Maxwell, 2004; Ioannidis, 2005), jednak jest standardowo cytowany przy okazji najnowszych dyskusji dotyczących „kryzysu”.

Artykuł ma dwie części: „empiryczną” i matematyczną. W części „empirycznej” przedstawiono sprawozdanie z rzeczywistego badania. Wybiórczo raportując dane, „dowiedziono” występowania absurdalnych efektów, np. słuchanie piosenki *When I'm 64* zespołu „The Beatles” *zmniejsza* wiek osób badanych. W części matematycznej wygenerowano losowe rozkłady danych, mające symulować rzeczywiste badania. Zastosowano jedną z czterech praktyk „podkręcania” danych, żeby sprawdzić, w jaki sposób zwiększą prawdopodobieństwo wyniku istotnego. Te praktyki to:

1. Używanie wielu zmiennych zależnych i raportowanie tylko tej, która „wyszła”.
2. Dodawanie kolejnej grupy badanych tak długo, aż nie osiągnie się istotności statystycznej (i, oczywiście, zaprzestawanie badań, gdy tylko istotność zostanie osiągnięta).
3. Uwzględnianie jakiejś zmiennej dwuwartościowej jako dodatkowej współzmiennej (np. płci).
4. Wykonywanie eksperymentu z więcej niż dwoma grupami eksperymentalnymi i wybiórcze raportowanie tylko tych, między którymi wykryto istotne różnice.

Stosowanie wymienionych praktyk zwiększa około dwukrotnie prawdopodobieństwo popełnienia błędu fałszywie pozytywnego z „tradycyjnych” 5%, jednak stosowanie wszystkich czterech metod powoduje, że istnieje aż 61% szans na znalezienie jakiegoś istotnego efektu w zupełnie losowych danych.

Fakt, że teoretycznie jest możliwe stosowanie takich sztuczek, nie mówi nam nic na temat tego, jak bardzo jest to rozpowszechnione. Dlatego próbę empirycz-

nego oszacowania rozpowszechnienia wyżej wymienionych wątpliwych praktyk badawczych przeprowadzili John i współautorzy (2012). Zaproszenie do badania wysłano do 6 tys. psychologów, pracowników naukowych amerykańskich uczelni. Badanym zapewniono anonimowość i motywowano do mówienia prawdy (wskazywali organizacje pożytku publicznego, na których konto dokonywano przelewów). Ostatecznie nieco ponad 2100 osób wzięło w nich udział. Badacze pytali o to, czy osoba kiedykolwiek realizowała jedną z wymienionych w Tabeli 1 wątpliwych praktyk. Dodatkowo ankietowani badacze szacowali rozpowszechnienie tych praktyk wśród innych badaczy oraz oceniali, jaka część badaczy przyznałaby się do wątpliwych praktyk i zestawiając te wartości, szacowano rzeczywiste rozpowszechnienie praktyk. Badani byli motywowani do mówienia prawdy (autorzy wpłacali pewną kwotę na wybraną organizację dobroczynną). Pytano także o dopuszczalność poszczególnych praktyk w 3-stopniowej skali. Wyniki przedstawiłem w Tabeli 1.

W przypadku niektórych kategorii ponad 50% badanych przyznaje się do praktyk, które prowadzą do zawyżonej liczby wyników fałszywych pozytywnych. Oszacowano, że odsetek badaczy stosujących analizowane przez Simmonsa i współautorzy (2011) praktyki, takie jak nieraportowanie zmiennych zależnych, dokładanie osób do celek aż do skutku i wybieranie tylko „działających" porównań między grupami, wynosi odpowiednio 78%, 72% i 42%. Jednym z najbardziej pesymistycznych wniosków jest, że wszystkie wątpliwe praktyki badawcze (poza fałszowaniem danych) nie są postrzegane jako coś złego, ale są oceniane jako „dopuszczalne" (rzadziej jako „raczej dopuszczalne"). W przyszłych badaniach byłoby interesujące rozstrzygnąć, czy wynika to z niskiej świadomości metodologicznej, czy też raczej z cynicznego pragmatyzmu. Można też zasadnie zakładać, że jakiś odsetek nieuczciwych badaczy nie przyzna się do stosowania takich praktyk lub w ogóle nie przystąpi do ankiety (nie wzięło w niej udziału ponad 60% zaproszonych).

Wyniki Johna, Lowensteina i współautorów wskazują, że „psychologia fałszywie pozytywna", przed którą ostrzegają Simmons i współautorzy (2011), może mieć miejsce, i nie jest tylko matematyczną ciekawostką. Idąc krok dalej, Bakker, van Dijk i Wicherts (2012) symulowali badania nad efektami o założonej sile i zestawiali je w „metaanalizy". Najlepszą strategią pod kątem uzyskania publikowalnych (istotnych statystycznie) wyników było przeprowadzanie wielu niedużych badań z zastosowaniem wątpliwych praktyk badawczych. Takie praktyki najbardziej też wypaczały obraz rzeczywistości. Biorąc pod uwagę, że standardowa wielkość badania psychologicznego to około 40 osób (Tressoldi, 2012;

Marszalek, Barber, Kohlhart i Holmes, 2011), wydaje się, że taka strategia jest nierzadko stosowana (por. też Francis, 2014).

Tabela 1

Wątpliwa praktyka badawcza	Procent badanych, którzy przyznali się do stosowania danej praktyki	Szacowane przez ankietowanych badaczy rozpowszechnienie danej praktyki w społeczności naukowej	Rozpowszechnienie danej praktyki wg autorów badania	Średnia ocena dopuszczalności danej praktyki (w nawiasie odchylenia standardowe)
1. Nieraportowanie wszystkich wykorzystanych zmiennych	66,5%	60%	78%	1,84 (0,39)
2. Zbieranie dodatkowych danych po sprawdzeniu, czy już posiadane dane są istotne	58,0%	62%	72%	1,79 (0,44)
3. Wybiórcze raportowanie tylko tych warunków eksperymentalnych, pomiędzy którymi zanotowano istotne różnice	27,4%	38%	42%	1,77 (0,49)
4. Rezygnacja ze zbierania danych wcześniej niż zaplanowano, ze względu na znalezienie oczekiwanych wyników	22,5%	41%	36%	1,76 (0,48)
5. Nieuprawnione zaokrąglanie wartości p (np. raportowanie wartości $p = 0{,}054$ jako $p < 0{,}05$)	23,3%	41%	39%	1,68 (0,57)
6. Selektywne raportowanie tylko tych badań, które „wyszły”	50,0%	61%	67%	1,66 (0,53)
7. Decydowanie o tym, czy wykluczyć określone dane po sprawdzeniu wpływu takiej operacji na rezultaty	43,4%	45%	62%	1,61 (0,59)
8. Opisywanie nieoczekiwanego wcześniej odkrycia jako przewidzianego od samego początku	27,0%	50%	54%	1,50 (0,60)
9. Twierdzenie, że na wyniki nie mają wpływu zmienne demograficzne, podczas gdy w rzeczywistości nie wiadomo lub mają wpływ	4,5%	22%	13%	1,32 (0,60)
10. Fałszowanie (fabrykowanie) danych	1,7%	10%	9%	0,16 (0,38)

Źródło: dane zawarte w John i współautorzy (2012). *Uwagi.* W kolumnie 1 przedstawiono wyniki grupy, która była dodatkowo motywowana do mówienia prawdy. Wyniki grupy kontrolnej były zwykle kilka procent niższe (tj. niższe o 0-7%). Wartość w kolumnie 2 szacunkowo na podstawie wysokości słupka wykresu (brak dokładnych danych surowych w artykule). Odsetek w kolumnie 3 jest to oszacowanie rozpowszechnienia danej praktyki przez autorów badania na podstawie odsetka przyznających się i ocenianego przez badanych odsetka osób, które przyznałyby się. Użyta w kolumnie 4 skala: 0 – całkowicie niedopuszczalne; 1 – w pewnym stopniu dopuszczalne; 2 – dopuszczalne.

Wydaje się, że w psychologii złożoność przedmiotu badania, niski stopień konsensusu co do sposobu mierzenia zmiennych psychologicznych i płynność założeń teoretycznych ułatwiają badaczom podejmowanie arbitralnych decyzji przy analizie i prezentacji danych. Aby lepiej zilustrować ten problem, podam przykład hipotetycznego badania zaczerpnięty od Gelmana i Lokena (2013) i twórczo przeze mnie rozwinięty. Badanie to dotyczy różnic w rozwiązywaniu problemów matematycznych między zwolennikami demokratów i republikanów w Stanach Zjednoczonych. Badacz zakłada, że różnice występują w zależności od kontekstu, w jaki „ubierze się" problem: demokraci lepiej poradzą sobie z problemem, który dotyczy opieki zdrowotnej, a republikanie lepiej z matematycznie analogicznym problemem, ale „ubranym" w kontekst działań militarnych. Badacz zbiera też szereg dodatkowych danych demograficznych. Załóżmy, iż okazuje się, że efekt ten występuje tylko u mężczyzn. Można to łatwo wytłumaczyć, wszak mężczyźni mają silniejsze przekonania ideologiczne (tu hipotetyczny badacz może przywołać szereg danych). A co jeśli różnice dotyczą tylko kobiet? Badacz też stworzy do tego bogate uzasadnienie wskazując na to, że kobiety są bardziej wrażliwe na kontekst (i wzmocni to szeregiem źródeł). Idąc dalej, wiemy, że badacz pytał o sympatie partyjne na skali 1-7. Powstaje kolejne pytanie, kogo ma porównywać? Osoby, które zaznaczyły 1-3, z osobami 5-7? Może lepiej 3 i 5 odrzucić jako osoby bliskie neutralności? A może porównywać ideologiczne „jedynki" i „siódemki"? Co zrobić z osobami neutralnymi? Wyłączyć je, a może kluczowe porównania wskażą na różnice między osobami zaangażowanymi politycznie a neutralnymi? Może żadna analiza nie da istotnego efektu, ale jeśli „wrzucimy" wiek do modelu, to okaże się, że występuje istotna interakcja, do której można dorobić dalszą historię (np. młodsi są bardziej skrajni w swoich postawach politycznych vs młodsi mają mniej skrystalizowane i bardziej zmienne poglądy). Po drodze nasz hipotetyczny badacz może jeszcze podjąć nieskończoną liczbę innych arbitralnych decyzji, jeśli uzna, że dają istotne efekty.

Na dowolnym dużym zbiorze danych surowych można wykonać ogromną liczbę analiz, z których jakaś część będzie istotna statystycznie przez sam przypadek. Dodatkowo do każdego wyniku jesteśmy w stanie stworzyć lepsze lub gorsze uzasadnienie (czy mówiąc dosadnie – „historyjkę") i z nieprzebranej literatury znaleźć jakieś źródło, które ją wspiera. Artykuły Stapela zawierają dziesiątki cytowań uzasadniających hipotezy do jego badań, choć wiemy, że żadnych badań nie było. Już po poznaniu wyników badacz może je przedstawić jako przewidziane od samego początku przez swoje „teorie", i w ten sposób nadać tym teoriom rzekomej, ale nieuprawnionej skuteczności w generowaniu empi-

rycznie sprawdzonych przewidywań. Niestety, jest to praktyka wcale nierzadka (Kerr, 1998).

Gelman i Loken (2013) analizują kilka przykładów rzeczywistych badań, wskazując na trudną do wytłumaczenia arbitralność niektórych decyzji. Tutaj pozwolę sobie przedstawić inne badanie, które budzi moje wątpliwości z tego samego względu. Gervais i Norenzayan (2012) przedstawili badania wskazujące na to, że analityczne myślenie jest predyktorem braku wiary religijnej. W czterech eksperymentach manipulacja polegała m.in. na oglądaniu rzeźby myślącego filozofa (lub w grupie kontrolnej – rzeźby atlety), prymowaniu słowami związanymi z analitycznym myśleniem lub neutralnymi, zmuszaniu do czytania słów z niewyraźną czcionką i wyraźną itp. W każdym z badań uzyskano wyniki wskazujące na związek „analitycznego myślenia” (ściślej: manipulacji wzbudzającej rzekomo takie myślenie) z brakiem wiary. Wątpliwości budzi jedna rzecz: w poszczególnych badaniach eksperymentalnych używano innych zmiennych zależnych. I tak w pierwszym eksperymencie badani określali poziom wiary w Boga na skali od 0 do 100; w drugim odpowiadali na pytania dotyczące ich wiary w Boga, anioły i diabła, każde na skali od 1 do 7; a w kolejnym ustosunkowywali się do 10 pytań na temat ich religijności (każde w skali od 1 do 7; np. „Moje przekonania religijne są tym, co naprawdę leży u podstaw mojej filozofii życiowej”). Jednak nawet w tej niekonsekwencji badacze byli *niekonsekwentni*, gdyż w ostatnim badaniu wrócono do zmiennej zależnej „wiara w Boga w skali 0-100”. Autorzy nie uzasadnili, dlaczego używali różnych zmiennych zależnych w każdym badaniu. O ile różnorodne procedury eksperymentalne są jeszcze jakoś zrozumiałe, o tyle różnorodne zmienne zależne utrudniają rzetelną interpretację wyników, ponieważ można się zastanawiać, czy poszczególne zmienne zależne mierzą dokładnie to samo. Przypuszczam intuicyjnie, że miary te prawdopodobnie mierzą coś bardzo zbliżonego, tym niemniej ich arbitralne stosowanie może wskazywać na to, że każdorazowo używano kilku zmiennych zależnych i przytoczono tylko te, w przypadku których uzyskano istotne zależności (lub były różne kombinacje manipulacji i zmiennej zależnej i przedstawiono tylko działające). Na wybiórcze raportowanie badań wskazuje bardzo wysoka negatywna korelacja między siłą efektu a wielkością próby. W czterech badaniach eksperymentalnych korelacja ta wynosi: r = -0,97. Matematycznie patrząc, sama wielkość próby nie ma związku ze średnią wielkością efektu. Ma jednak związek z wariancją efektów (por. Lippa, 2009), tj. w przypadku mniejszych prób efekty będą bardziej zróżnicowane. Przy mniejszych próbach siłą rzeczy tylko efekty silniejsze „załapią” się na istotność. Tak wysoka korelacja pomiędzy siłą efektu a wielkością próby może być wynikiem przypadku, ale może też

wskazywać na to, że istniały dodatkowe badania, o których badacze nie wspomnieli (por. szczególnie Francis, 2012). W podobnym duchu Fergusson (2013) pokazał, jak naukowcy, badający wpływ gier komputerowych na agresję, używając zbliżonej procedury rażenia innych białym szumem, wyciągali z niej różne wskaźniki, zapewne w zależności od tego, które przynosiły istotne zależności (np. liczba sesji, w których badani zaaplikowali najgłośniejsze szumy, pierwiastek kwadratowy z długości szumów, pomnożony przez intensywność, sumy intensywności albo średnie z intensywności i długości osobno itp.).

Zarówno fabrykacja, jak i wybiórcza prezentacja wyników („psychologia fałszywie pozytywna") są rodzajami nierzetelności, moim zdaniem w praktycznych skutkach niewiele się różniącymi. W jednym i drugim przypadku społeczność badaczy dostaje zniekształcony obraz rzeczywistości, otrzymuje informacje tylko o udanych badaniach, wskazujących na rzekomą łatwość uzyskania efektu, nie może wyliczyć rzeczywistej siły efektów, a nieopublikowane dane mogą zawierać informacje o jakichś ważnych moderatach niwelujących efekt. Postawię tezę, że o ile zarówno badacze, jak i redaktorzy czasopism zdają sobie sprawę z niegodziwości fałszerstwa, o tyle wybiórcze publikowanie jest rodzajem strategii wynikającej częściowo ze statystycznej niewiedzy (zob. akceptacja w badaniach John i in., 2012) oraz niepisanej umowy między redakcjami i badaczami, która zakłada, że atrakcyjność manuskryptu jest warunkiem nie mniej ważnym niż rzetelność. Publikowane będą więc tylko dane wspierające hipotezy, co istotnie spotykamy w czasopismach empirycznych. Według analizy bibliometrycznej Fanelliego (2010) psychologia wespół z psychiatrią jest nauką z największą liczbą artykułów, w których pozytywnie zweryfikowano wyjściową hipotezę. Nie wydaje się, żeby teorie psychologiczne były tak wysoce wyrafinowane, a narzędzia precyzyjne, aby psycholodzy lepiej przewidywali rzeczywistość niż przyrodnicy.

Pojawiło się w ostatnim czasie też wiele innych sygnałów świadczących o tym, że rzetelność danych w psychologii daleka jest od doskonałości. Szczególnie wymowne są analizy wskazujące na: niską moc badań psychologicznych (Bakker i in., 2012; Francis, 2012, 2014; Olechowski, 2012; Tressoldi, 2012), co przy niewielkiej liczbie publikowanych wyników negatywnych (Fanelli, 2010) wskazuje na mocno wybiórcze publikowanie (*publication bias*); rozkład wartości *p* w literaturze, zwłaszcza nieprawdopodobny wzrost częstości tych wartości nieco poniżej 0,05 (Masicampo i Lalande, 2012; Leggett, Thomas, Loetscher i Nicholls, 2013; Simonsohn i in., 2014); nieprawidłowości przy raportowaniu wartości *p*, szczególnie poprzez klasyfikowanie wartości nieistotnych jako istotne (Bakker i Wicherts, 2011; Wicherts i in., 2011); niewielkiej liczby replikacji,

zwłaszcza dokładnych (Makel, Plucker i Hegarty, 2012); niedzielenia się danymi surowymi (Mitchell, 2012; Wicherts, Borsboom, Kats i Molenaar, 2006).

UWAGI KOŃCOWE

Efektem „kryzysu” są szerokie dyskusje nad koniecznością zmian w praktykach badawczych i publikacyjnych w psychologii. Niektóre postulaty mają charakter bardziej techniczny, inne zakładają fundamentalną zmianę podejścia do uprawiania nauki. Wśród tych pierwszych znajduje się postulat obowiązkowej publikacji danych surowych (Wicherts, Bakker, 2012; Simmonsohn, 2013), co mogłoby ułatwić wykrywanie nietypowych wzorów danych (jak już wspominałem, w tekstach Stapela wykryto bardzo wiele błędów). Analiza danych pozwoliła Uri Simmonsohnowi (2013) wykryć kolejne dwa przypadki fabrykacji danych przez psychologów społecznych (konkretnie przez Lawrence’a Sanna i Dirka Smeestersa). W innej niedawnej historii statystycy wskazali skrajnie niskie prawdopodobieństwo danych uzyskanych przez Jensa Förstera (Kolfshooten, 2014). Łatwość dostępu do danych surowych może sprzyjać wykrywaniu oszustw, choć oczywiście nie uniemożliwi ich całkowicie. Inny postulat to obowiązkowa prejestracja badań w celu ograniczenia możliwości „twórczego” wyciągania prawidłowości z ogromu danych (Aveyard i in., 2013). Ten postulat nie powinien być trudny do zrealizowania, gdyż badania przed wykonaniem są zwykle zgłaszane do komisji etycznych, dodatkowa rejestracja ich w dedykowanych portalach nie wydaje się dużym wysiłkiem. Inny pomysł na ograniczenie „psychologii fałszywie pozytywnej” to obowiązkowa deklaracja przez autorów, czy przedstawili w raporcie wszystkie zmienne, warunki, sposoby ustalania wielkości próby oraz usuwania przypadków odstających (Simmons i in., 2012).

Istnieją wreszcie postulaty, które zakładają bardziej fundamentalne reformy. Czasopisma są otwarte przede wszystkim na wyniki pozytywne, natomiast nawet najlepsze badania z nieistotnymi wynikami zwykle mają drogę zamkniętą do publikacji. Jeśli badanie nie przynosi oczekiwanych rezultatów (jak to się na ogół dzieje w nauce) badacze mają do wyboru zmarnowanie kilku miesięcy pracy (a zatem mniejsze szanse na etat, awans itd.) albo dokonanie mniejszych i większych przekłamań (ewentualnie, jeśli przeprowadzili kilka eksperymentów, z których część tylko przyniosła oczekiwane wyniki, mogą pominąć te nieudane dla zwiększenia atrakcyjności narracji). Kluczowa wydaje się zatem zmiana systemu zachęt, tak żeby promować dobre badania, niezależnie od wyników (Nosek i Bar-Anan, 2012). Ciekawe rozwiązanie zaproponowało czasopismo *Cortex*,

które zadeklarowało otwarcie sekcji prejestrowanych badań (Chambers, 2013). Autorzy mają wysyłać artykuły z sekcją wstępną oraz metodą. Recenzenci mają ocenić, czy postawione pytanie badawcze jest ważne, a badanie sensownie zaprojektowane. W przypadku pozytywnej decyzji autor dopiero wtedy robi badanie, a czasopismo zobowiązuje się wydrukować artykuł niezależnie od wyników. W ten sposób autorzy mogą potencjalnie zmniejszyć ryzyko straty czasu i zasobów na „niepublikowalne" badania. Jeśli natomiast dane są już zebrane, zaproponowano, aby część recenzentów oceniała teksty bez znajomości wyników (*result blind review*; Greve, Bröder i Erdfelder, 2013), a ewentualne rozbieżności mogłyby być dla redaktorów wskazówką, że dobre badanie traci przez negatywne wyniki (lub wątpliwe badanie zyskuje na wynikach pozytywnych). Ostatecznym kryterium oceny danych w nauce jest oczywiście możliwość ich niezależnego potwierdzenia, dlatego duża część głosów dotyczy zwiększenia roli replikacji dokładnych w psychologii (Asendorpf i in., 2013; Nosek, Spies i Motyl, 2012; Francis, 2012).

Te propozycje nie wymagają dużych nakładów finansowych. Wymagają jednak przełamania zwyczajów i nawyków obecnych w dyscyplinie od dziesiątków lat. Większość najważniejszych graczy nie podjęła jeszcze decyzji o zasadniczej zmianie reguł gry. Czasopisma, takie jak *Journal of Personality and Social Psychology* czy *Journal of Experimental Psychology*, które najbardziej ucierpiały na oszustwie Stapela, powinny być najbardziej zainteresowane zmianami. Zbyt wiele nieprawidłowości zostało wykazanych w ostatnich latach i wydaje się, że prędzej czy później praktyki badawcze i publikacyjne będą musiały ulec zmianie.

Zaryzykuję na koniec tezę, że praźródłem obecnego kryzysu jest pogoń za „nowatorskością". Powoduje to skakanie od hipotezy do hipotezy bez dogłębnego zrozumienia jednego efektu i uzyskania na jego temat maksymalnie pewnej wiedzy (na tyle, na ile jest to możliwe w nauce takiej, jak psychologia). Najlepsze czasopisma wymagają „odkrywczych" badań, i podobne wymagania stawiają instytucje przyznające granty (oto przykład z naszego podwórka: o programie Maestro na stronie NCN przeczytamy: „jest to konkurs na finansowanie projektów badawczych mających na celu realizację pionierskich badań naukowych, [...] wykraczających poza dotychchzasowy stan wiedzy"). Duże, nawet najlepiej pomyślane badanie mające skontrolować poprawność innych badań nie dostanie zatem grantu. Sygnał jest jasny: wybitny naukowiec nie angażuje się w replikacje dokładne, w najlepszym wypadku może powtórzyć badanie, twórczo je modyfikując (tzw. replikacje konceptualne). Jednak sama replikacja konceptualna zasadniczo niewiele mówi o oryginalnym efekcie, w szczególności nie wiadomo, czy niemożność powtórzenia wynika z niewystępowania efektu, czy z wprowa-

dzonych modyfikacji. Koniec końców, odkrycia się kumulują, nie mamy niezawodnych kryteriów odróżniania prawdziwych od fałszywych, a ze względów tylko logistycznych nie będzie nigdy możliwe zreplikowanie wszystkich badań (Makel i in., 2012). Jakaś część przekłamań, wynikających z arbitralnych decyzji, fałszerstw albo nieuświadamianych błędów, nie zostanie nigdy skorygowana. Nie będąc chemikami ani fizykami, nie mamy możliwości uzyskiwania łatwo powtarzalnych i w miarę teoretycznie zrozumiałych wyników (choć i oni nie zawsze mają takie możliwości). Nierealistyczne jest założenie, że w takiej nauce, jak psychologia kilka badań pod rząd musi zawsze pokazywać ten sam efekt (zwłaszcza jeśli efekt ten jest słaby, co też kłóci się z elementarnym rachunkiem prawdopodobieństwa; Francis, 2012, 2014). Podobnie nierealistyczne jest założenie, że czasopisma będą wypełnione od pierwszej do ostatniej strony „odkryciami”. Sam cenię artykuły opisujące nowe idee, które zwiększają nasze rozumienie rzeczywistości i przełamują stare schematy. Ale nauka taka jak psychologia w równie dużym stopniu musi się opierać na żmudnej, mało kreatywnej pracy polegającej na sprawdzaniu prawdziwości wiedzy wypracowanej przez innych. Nie sądzę, że psychologia straciłaby, gdyby obok nowatorskich, eksploracyjnych i względnie niedużych badań na łamach czasopism były też promowane teksty z odtwórczymi badaniami, ale o bardzo dużej mocy i zrobionymi w sposób maksymalnie rzetelny.

LITERATURA CYTOWANA

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence, 149*(1), 91-130.

Asch, S. E. (1952/1987). *Social psychology*. New York: Oxford University Press.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., […] i Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology: Recommendations for increasing replicability. *European Journal of Personality, 27*(2), 108-119.

Aveyard, P., Bellgrove, M., Bertamini, M., Bestmann, S., Bishop, D., Brembs, B., ... Wolfe, J. (2013). *Trust in science would be improved by study pre-registration.* Strona internetowa: http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Bakker, M., van Dijk, A. i Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Bakker, M. i Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Bevan, W. (1991). Contemporary psychology: A tour inside the onion. *American Psychologist, 46*, 475-483.

Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times*. Strona internetowa: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?_r =0

Brzeziński, J. M. (2012). Co to znaczy, że wyniki przeprowadzonych przez psychologów badań naukowych poddawane są analizie statystycznej? *Roczniki Psychologiczne, 15*(3), 7-39.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610.

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist, 30*(10), 967-976.

European Association of Social Psychology (2012). *Statement EASP on Levelt December 2012*. Strona internetowa: http://www.easp.eu/news/Statement%20EASP%20on%20Levelt_December_%202012.pdf

Fanelli, D. (2010). „Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*, e10068.

Ferguson, C. J. (2013). Violent video games and the Supreme Court: Lessons for the scientific community in the wake of Brown v. Entertainment Merchants Association. *American Psychologist, 68*(2), 57-74.

Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Scienc*e, *7*(6), 585-594.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*, 1180-1187.

Gelman, A. i Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Strona internetowa: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gervais, W. M. i Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*(6080), 493-496.

Greve, W., Bröder, A. i Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist, 18*(4), 286-294.

Ioanis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

John, L. K., Loewenstein, G. i Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kepes, S. i McDaniel, M. A. (2013). How trustworthy is the scientific literature in Industrial and Organizational Psychology? *Industrial and Organizational Psychology, 6*(3), 252-268.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Klebaniuk, J. (2012). Profesor Stapel na dopingu. O upiększaniu psychologii społecznej. *Psychologia Społeczna, 7*, 213-217.

Kolfshooten, F. (2014). Fresh misconduct charges hit dutch social psychology. *Science, 344*, 566-567.

LeBel, E. P. i Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371-379.

Leggett, N. C., Thomas, N. A., Loetscher, T. i Nicholls, M. R. (2013). The life of p: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology, 66*(12), 2303-2309.

Levelt Committee, Noort Committee, Drenth Committee (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Strona internetowa: https://www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2013/01/finalreportLevelt1.pdf

Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior, 38*(5), 631-651.

Makel, M. C., Plucker, J. A. i Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542.

Marszalek, J. M., Barber, C., Kohlhart, J. i Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills, 112*, 331-348.

Masicampo, E. J. i Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163.

Mitchell, G. (2012). What is wrong with social psychology? *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology, 26*(2), 12-17.

Murayama, K., Pekrun, R. i Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*, 107-118.

Nosek, B. A. i Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*, 217-243.

Nosek, B. A., Spies, J. R. i Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Olechowski, M. (2012). Kryzys psychologii, psychologia kryzysu. *Psychologia Społeczna, 22*, 227-233.

Pashlev, H. i Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Rand, K. i Ilardi, S. (2005). Toward a consilient science of psychology. *Journal of Clinical Psychology, 61*, 7-20.

Reich, E. S. (2009). *Plastic fantastic: How the biggest fraud in physics shook the scientific world*. New York: Palgrave Macmillan.

Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology, 3*, 113-123.

Rozin, P. (2001). Social psychology and science: Some lessons from Salomon Asch. *Personality and Social Psychology Review, 5*, 2-14.

Simmons, J. P., Nelson, L. D. i Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simmons, J. P., Nelson, L. D. i Simonsohn, U. (2012). A 21-word solution. *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology, 26*, 4-7.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875-1888.

Simonsohn, U., Nelson, L. D. i Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Staats, A. (1999). Unifying psychology requires new infrastructure, theory, method, and a research agenda. *Review of General Psychology, 3*, 3-13.

Stapel, D. A. i Lindenberg, S. (2011). Coping with chaos: How disordered contexts promote stereotyping and discrimination. *Science, 332*, 251-252.

Stroebe, W., Postmes, T. i Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*(6), 670-688.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology, 3*(218), doi: 10.3389/fpsyg.2012.00218.

Wicherts, J. M. i Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40*(2), 73-76.

Wicherts, J. M., Bakker, M. i Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE, 6*(11), e26828.

Wicherts, J. M., Borsboom, D., Kats, J. i Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726-728.